Genetic regulation of DNA methylation across tissues reveals thousands of molecular links to complex traits

Meritxell Oliva (✉ meritxellop@gmail.com)
University of Chicago https://orcid.org/0000-0002-5068-213X

Kathryn Demanelis
University of Pittsburgh

Farzana Jasmine
University of Chicago

Yihao Lu
University of Chicago

Habibul Hahsan
University of Chicago

Muhammad Kibriya
University of Chicago https://orcid.org/0000-0001-9784-6958

Lin Chen (✉ lchen@health.bsd.uchicago.edu)
University of Chicago

Brandon Pierce (✉ brandonpierce@uchicago.edu)
University of Chicago

Article

Keywords: DNA methylation, complex traits, epigenetic modifications

Posted Date: May 11th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-492596/v1

Version of Record: A version of this preprint was published at Nature Genetics on December 12th, 2022. See the published version at https://doi.org/10.1038/s41588-022-01248-z.

Abstract

Epigenetic modifications of chromosomes, including DNA methylation (DNAm), play a fundamental role in gene regulation in humans. We generated DNAm data for 987 samples from the Genotype-Tissue Expression (GTEx) project, representing 9 tissue types and 424 subjects. We integrated GTEx RNA-seq data to examine methylome-transcriptome associations, their tissue specificity, and their overlap with regulatory regions. We mapped DNAm quantitative trait loci in *cis* (mQTLs), contrasted mQTLs with expression QTLs (eQTLs) with respect to functional elements, and assessed their relative contributions to complex traits. We identified thousands of mQTL links to traits in locations lacking a relevant eQTL. By integrating genetic, diverse -omics and phenotype data, we contribute to the understanding of molecular regulatory mechanisms in human tissues and their association with complex traits.

Introduction

The majority of common genetic variants that impact human traits are believed to exert their effects through the regulation of nearby genes [1,2]. Genetic regulation of gene expression has been comprehensively characterized across many human tissue types [3,4], and expression quantitative trait loci (eQTLs) appear to underlie a substantial fraction of variant-trait associations [5–7]. However, our understanding of the regulatory mechanisms by which variants influence human traits is far from complete. Elucidating how variants impact epigenetic features, e.g. DNA methylation (DNAm) is critical, as these features can influence, and respond to, gene expression [8].

While gene expression data from healthy tissues and cells are widely available and well-characterized [9], comparable data sources for DNAm are lacking, and exist primarily for blood and for diseased tissues such as tumors [10]. The enhanced GTEx project [11] seeks to complement existing gene expression data from human tissues with additional molecular traits, including DNAm.

In humans, DNAm at CpG dinucleotides is involved in key biological processes [12–14] and plays a critical role in the etiology of numerous diseases [15]. Among other factors, inter-individual DNAm variation is influenced by genetic variation [16], and the integration of DNAm quantitative trait loci (mQTLs) with genome-wide association analysis (GWAS) data has uncovered a putative causal role of DNAm in genetic susceptibility mechanisms [17–20]. With notable exceptions [17–19,21–23], most mQTL studies conducted to date have assessed DNAm in blood-derived samples [20,22–26], preventing the identification of tissue-specific mQTLs undetectable in blood. Therefore, mQTL catalogs derived from a variety of healthy, solid tissues are needed and can contribute to the characterization of the etiology of complex traits.

Here, we profiled human DNAm for > 750,000 genomic locations (CpGs or CpG sites). The dataset includes 987 samples representing 9 tissue types (Supplementary Table 1). Gene expression is also available for 3,872 samples derived from these 9 tissue types [4], and 495 DNAm-profiled samples have gene expression and genotype data available. To further characterize the relationship between DNAm and gene expression in a tissue-specific manner, we identified CpGs for which DNAm was associated with local gene expression. To characterize the *cis*-genetic regulation of methylation across tissues and compare it to its gene expression counterpart, we mapped mQTLs and eQTLs. We contrasted mQTL to eQTL profiles and characterized their interdependency as well as differential tissue specificities, functional mechanisms and regulatory pleiotropy signatures. To evaluate the impact of mQTLs on human traits and assess their relative contribution compared to eQTLs, we integrated mQTLs and eQTLs with genome-wide association study (GWAS) summary statistics of 87 GWASs.

Results

Comparison of DNA methylomes with transcriptomes reveals tissue-specific signatures unique to methylation

To investigate the contribution of tissue type to the similarity of DNAm-profiled samples, we reduced the high-dimensional methylome dataset to two dimensions, which revealed sample similarity by tissue of origin (Fig. 1a). We identified robust methylation-derived tissue clusters (Methods) that reflect shared yet distinct patterns compared to transcriptome-derived clusters (Fig. 1b). Observed patterns suggest that the biological processes that drive methylome similarity across tissues differ, in part, from corresponding transcriptome ones, and that blood, despite being the most widely studied biospecimen type for DNAm, is not a suitable proxy of solid tissue DNAm patterns.

Local correlation between DNA methylation and expression can be tissue-specific or shared across tissues

To characterize the interdependence between DNAm and expression, we mapped DNAm associations with local gene expression (eQTMs) across tissue types. The number of CpGs correlated with at least one gene (eCpGs) ranged from 1,542 in testis to 4,568 in ovary (Supplementary Table 2). We assessed replication of eQTMs in a muscle-derived external dataset [19] by estimating the true positive rate (π1) of eQTMs identified in each GTEx tissue type. We observed a high replication rate (cross-tissue average π1 = 0.75), confirming the reproducibility of the eQTMs mapped herein despite limited samples available. Similarly to transcriptome- and eQTL-derived patterns observed in previous GTEx analyses [3,4,9], eQTMs tend to be either tissue-specific or shared across most tissue types (Supplementary Figure 1).

To investigate how accurately the CpG-gene assignments provided by Illumina reflect the eQTM results observed in GTEx data, we contrasted our eQTM findings with the EPIC array CpG annotation file provided by Illumina. We observed that only 45% (2,641/5,898) of detected eCpGs match an Illumina CpG-gene association (Supplementary Note). Thus, our eQTM predictions can be utilized to assign CpGs to gene(s) with which they are biologically linked.

The association of DNA methylation with expression of nearby genes differs by regulatory elements

To functionally characterize the relationship between CpGs and genes, we identified enrichments (Fisher's exact test FDR < 0.01) of eCpGs in gene regulatory elements compared to CpGs uncorrelated with gene expression. Promoters and proximal enhancers are strongly enriched (OR = 4.55 and 4.17, respectively) for eCpGs, as well as insulators (OR = 2.32) and distal enhancers (OR = 1.74). Overall, 54% (3,188/5,898) of eCpGs overlap with gene regulatory elements. Using logistic regression, we examined additional factors for evidence of contribution to eQTM presence (Supplementary Table 2), which revealed distinctive signatures by regulatory element class (Fig. 1c). Proximity of CpG to gene transcription start site (TSS) increases the likelihood of the eQTM association ubiquitously across regulatory elements, but high CpG methylation and corresponding negative correlation with gene expression are only predictive of eQTMs linked to promoters and proximal enhancers. Distal enhancer and insulator eQTMs are enriched for low-methylated eCpGs, and low gene expression appears to be predictive of insulator-linked eQTMs exclusively.

Together, these results suggest that DNAm in CpG sites is associated with transcription of *cis*-genes through their regulatory elements, as described [27]. However, they also indicate heterogeneity of the biological mechanisms driving eQTMs, that depend on the class of regulatory elements involved, compatible with patterns observed in blood cells [23,28].

Genetic regulation of DNA methylation in *cis* exhibits tissue specificity

To characterize the genetic regulation of DNAm across tissues, we mapped genetic variants that affect DNAm levels of proximal CpG sites in *cis* (mQTLs). We first analyzed each tissue separately by fitting a linear model that accounts for technical and biological factors related to DNAm variability, and identified significant (FDR < 0.05) mQTL CpG sites in *cis* (mCpGs). Subsequently, we modeled mQTL effects jointly across tissues to achieve increased QTL-mapping power by leveraging tissue-shared effects [29]. Additionally, we mapped secondary *cis*-QTL signals using a stepwise regression procedure [30]. To be able to compare mQTL to *cis*-eQTL patterns, we employed an analogous QTL-mapping approach to identify eQTLs.

We detected a total of 286,152 mCpGs, ranging from 108,844 in testis to 206,802 in lung (Fig. 2a), of which 45,543/286,152 (16%) have secondary signals in at least one tissue. For blood and muscle tissues, we quantified the mQTL replication rate in external datasets and observed a high replication rate (cross-tissue average π1 = 0.92). For a particular CpG, genetic regulation of DNAm tends to be either highly tissue-specific or highly shared across tissue types (Supplementary Figure 2), similarly to the observed pattern reported for eQTLs and splicing QTLs (sQTLs) [4]. On average, 37% of mCpGs observed in a single tissue are also present in all the remaining tissues (Fig. 2b), but only 5% of the identified mCpGs were detected as mCpGs exclusively in a single tissue. Compared to eQTLs, mQTLs appear to be significantly (Wilcoxon rank-sum test $P < 0.05$) more shared across tissues, as observed in blood cells [23].

Functional mechanisms that drive genetic regulation of DNA methylation differ from gene expression

To characterize differential molecular mechanisms of mQTLs relative to eQTLs, we integrated annotation of CpG islands (CGIs), genomic functional elements, and chromatin states (Supplementary Table 3) and performed within- and meta-tissue enrichment analyses (Methods). We observe that eQTLs are more strongly enriched in open chromatin sites than mQTLs (Supplementary Figure 3). While both eQTLs and mQTLs are enriched in gene regulatory regions (Fig. 2c), only eQTLs are enriched in CGIs and gene transcripts, particularly in splicing and untranslated exon regions (UTR), as previously described [4]. Conversely, mQTLs are depleted from CGIs and genes, as previously observed in blood [20,25], but are strongly enriched in distal enhancers. Compatible patterns are observed when analyzing tissue-specific extended chromatin state predictions matching mQTL tissue source (Supplementary Figure 3). DNAm QTLs also show enrichment in putative insulators (Fig. 2c). To further characterize additional mQTL-associated transcription factors (TFs), and to distinguish between mQTL and eQTL distinctive TF binding sites (TFBS) associations, we integrated empirical annotation of TFBSs corresponding to 339 TFs. Considering corresponding TFBSs, we identified 126 TFs significantly enriched (FDR < 0.01, Odds Ratio (OR) > 1) in eQTLs or mQTLs (Supplementary Table 3). We observed remarkably different TFBS enrichment profiles for mQTLs compared to eQTLs. The eQTL enrichments with the smallest p-values across tissues correspond to TFs involved in basal transcription, e.g. RNA Polymerase II genes. Conversely, mQTLs are enriched, among other TFBSs, in binding sites of steroid receptors, e.g. ESR1 and NR2F2, and other proteins known to be involved in 3D organization of the genome (Supplementary Table 3, Fig. 2d).

Altogether, these results suggest that mQTLs and eQTLs largely diverge in their underlying biological mechanisms, driven by mostly distinct sets of TFs. While eQTLs result from variants altering gene body and regulatory elements, mQTLs result in part from variants altering non-genic, distal regulatory elements, elements bound by proteins involved in chromatin spatial conformation and long-range interactions, including insulators. The location of mQTLs with respect to open chromatin and actively transcribed regions indicates that, compared to eQTLs, mQTLs are more likely to reside in proximal regulatory regions that appear inactive in the mQTL-mapping context analyzed herein.

Genetic co-regulation of DNA methylation and gene expression is not pervasive and exhibits heterogeneity across regulatory elements

Given their divergent genomic enrichment profiles, it is expected that mQTLs and eQTLs are driven, at least in part, by different causal variants. To quantify the extent of eQTL-mQTL pairs (e/mQTL) that share a predicted causal variant, we performed e/mQTL colocalization, and observed that the proportion of detectable mQTLs that show clear colocalization with a detectable eQTL is moderate (Fig. 2e), as only 21% of mQTL loci are suggestively colocalized (PP4 > 0.5) with at least one eQTL. Despite limitations in accurately estimating this fraction, our results indicate that a considerable fraction of mQTLs do not show clear associations with local gene expression in the context analyzed herein.

Among e/mQTL colocalized variants, the direction of the effect on methylation and expression is significantly (exact binomial test $P < 2.2e-16$) more often (53%) in the opposite direction, as previously observed [26]. However, we observe significant (test of equal proportions $P < 2.2e-16$) differences between regulatory regions; the proportion of mCpGs corresponding to opposite e/mQTL effects tends to be larger in eGene-matching promoters and proximal enhancers (61%)

than in distal enhancers (52%) and minority (39%) in insulators. These observations are in line with the view that hypomethylation in proximal gene regulatory regions is associated with active transcription.

Genetic regulation of DNA methylation is characterized by molecular regulatory pleiotropy

In order to characterize the molecular pleiotropic nature of mQTLs, we quantified the number of mCpGs and eGenes involved in loci harboring mQTL-eQTL colocalizations (Methods). Overall, we observe pervasive pleiotropy; the majority (78%) of colocalized eQTLs-mQTLs impact multiple mCpGs, and a considerable minority (28%) impact multiple eGenes (Supplementary Figure 4). The largest pleiotropic set, identified in ovary, is led by variant rs6433571 and involves 114 mCpGs and 8 eGenes in the HOXD gene cluster region (Fig. 3a), associated with epithelial ovarian cancer [31]. This pleiotropic effect is not driven by ovary-specific gene expression (Fig. 3b) but by ovary-specific genetic regulation of DNAm and expression (Fig. 3, c and d). By means of QTL-GWAS colocalization, we identified 112/114 mCpGs and 7/8 eGenes as significantly (PP4 > 0.5) colocalized with ovarian cancer risk (Supplementary Table 4), including the HOXD1 and HOXD3 genes which have a suspected role in genetic risk of ovarian cancer [32,33], as well as less characterized genes, e.g. non-coding gene HAGLR. Together, these findings provide an illustrative example of how genetic variants can drastically modify the DNAm landscape in a long-range and tissue-specific manner, altering gene expression and impacting disease risk.

Genetic regulation of DNA methylation impacts complex trait associations extensively

Prior studies [17–20] have provided evidence that mQTLs can be associated with human phenotypes. To evaluate the impact of mQTLs on traits in a systematic manner, and compare their effects to those of eQTLs, we integrated QTLs with genome-wide association study (GWAS) summary statistics of 87 GWASs, 83 of which had at least one QTL-overlapping GWAS hit ($P < 5e\text{-}08$), and were therefore tested for colocalization with a robust multi-method approach (Supplementary Figure 5).

Across all GWASs, tissues, and QTL types, we identified a total of N = 12,922 significant (RCP > 0.3 and PP4 > 0.3) QTL-GWAS colocalizations - named simply 'colocalizations' (Supplementary Table 5). We observed that mQTL colocalizations were more abundant than eQTL colocalizations for almost all (91%) of GWASs (Fig. 4). The overlap between eQTL- and mQTL-GWAS colocalizations is moderate, with 27% (749/2,734) of GWAS hits colocalizing with both QTL types (e/mQTL-shared), 55% of hits colocalizing with at least one mQTL but with no eQTLs (mQTL-specific), and 18% of hits colocalizing with at least one eQTL but with no mQTLs (eQTL-specific).

These results highlight the importance of integrating different types of -omics data from different tissue sources, and considering secondary QTL signals, to maximize the expectation of identifying molecular links to inheritable traits.

Genetic regulation of DNA methylation facilitates the fine-mapping of trait-associated causal variants and characterization of regulatory mechanisms

Among mQTL-specific colocalizations, we identified an ovary-specific mQTL association (rs2853669-cg07380026, P = 6.7e-13) colocalized (PP4 = 0.84) with a breast cancer GWAS signal in the TERT locus (Fig. 5a). TERT expression is mostly undetectable in adult tissues but high in tumors, and the locus harbors multiple independent variants associated with several types of cancer risk [34]. Another example of a mQTL-specific colocalization is an ovary-specific, secondary mQTL association (rs7161194-cg05029961, P = 8.0e-15) that colocalized (PP4 = 0.98) with a body mass index GWAS signal in the microRNA-rich MEG9 locus (Fig. 5b), for which colocalization could not be identified considering the primary mQTL signal. The mVariant rs7161194 affects *cis*-microRNAs' expression [35].

For 19% (144/749) of the colocalized e/mQTL-GWAS shared loci, the mQTL-GWAS association shows greater colocalization probability than the eQTL-GWAS association in at least one tissue and/or independent QTL colocalization. We observe cases where the additional resolution to define potentially causal variants brought by mQTLs is small due to almost perfect linkage disequilibrium (LD) between lead mQTL and eQTL variants, as in the breast cancer linked NTN4 locus (Fig. 5c). However, in multiple instances the lead e/mQTL variants are in moderate or low LD, and the mQTL mirrors the GWAS association substantially more optimally than eQTL does, as in the hypertension-associated MYO9B locus (Fig. 5d). We also observe cases where the GWAS locus harbors association signals compatible with the existence of multiple independent causal variants, where GWAS colocalization with eQTLs and mQTLs contributes to differentiate and characterize these independent causal variants by their distinct colocalization patterns, as in the EFEMP2 locus linked to asthma (Fig. 5e). Together, these results provide evidence that integration of e/mQTL-GWAS colocalization signals can aid the fine-mapping of the causal variant(s) and better characterize the molecular mechanisms underlying complex traits, as shown [2,36,37].

Trait-linked methylation quantitative trait loci exhibit molecular regulatory pleiotropy and enrichment in trait-relevant tissues

Identification of QTL associations with traits in relevant tissue(s) can provide insights into their underlying genetic and molecular mechanisms [6]. By analyzing the observed proportions of mQTL-GWAS colocalizations per tissue, we identified 18/65 traits with a disproportionate (test of equal proportions FDR < 0.05) amount of colocalizations in at least one tissue (Fig. 6a). Overall, the tissue with the largest proportion of colocalizations per trait matched the prior given current biological knowledge. For instance, blood clot and cell count traits were enriched in colocalizations derived from whole blood mQTLs, and breast cancer was nominally (Fisher's exact test P = 0.02) enriched in breast-derived mQTLs. For traits where the observed tissue link is less obvious, the observed enrichment can be artifactual or it could point to an uncharacterized role of a specific tissue in the trait's biology. Together, these results suggest that mQTLs are informative of complex traits' relevant tissues and can thereby aid the characterization of trait etiology, as observed for eQTLs [5–7] .

It is expected that many QTLs that impact traits do so by exhibiting molecular regulatory pleiotropy, i.e. altering multiple molecules and/or molecular phenotypes. It has been shown that eQTLs where the lead eVariant regulates multiple eGenes are more likely to yield a trait association than eQTLs that regulate a single eGene [4]. However, the effect that regulatory pleiotropy plays in mQTL-trait associations has not been extensively characterized. Here, we observe that mQTL-GWAS colocalizations are enriched in mVariants regulating multiple, as opposed to single, mCpGs (OR = 2.65, Fisher's exact test P < 2.2e-16). Among trait-linked mCpGs that colocalize with at least one eGene, we observed enrichment (OR = 1.40, P = 6.3e-08) for trait colocalizations involving multiple mCpGs and eGenes (Tier 4 in Supplementary Figure 3), as in HOXD locus (Fig. 3a). These results indicate that mQTLs that exhibit regulatory pleiotropy have increased chances to impact a complex trait.

Trait-linked genetically-regulated methylated loci exhibit decreased methylation and are preferentially located in regulatory regions

To better understand the DNAm changes that contribute to mQTL impactfulness on traits, we characterized DNAm levels of trait-linked mCpGs and their overlap with open chromatin regions, as well as DNAm gains and losses attributable to increased disease-risk alleles. Compared to mCpGs without identifiable trait links, we observe an enrichment (Fisher's exact test P < 2.2e-16, OR = 1.50) of trait-linked mCpGs in in open chromatin regions; with which 53% (1,791/3,381) of trait-linked mCpGs overlap, with e/mQTL-shared GWAS-colocalized loci exhibiting a stronger enrichment (OR = 1.96) compared to mQTL-specific loci (OR = 1.36). Trait-linked mCpGs tend to have lower DNAm levels compared to trait-agnostic mCpGs (Wilcoxon rank-sum test P < 0.05), whereas trait-linked eGenes tend to be highly expressed (Supplementary Figure 6). To characterize the pathogenicity of genetically-regulated DNAm changes, we examined the association of disease-risk alleles with the direction of mQTL effects for a subset of 31/87 disease GWAS traits, and we did not observe a global, significant (FDR < 0.05) association of disease-risk alleles with increased or decreased DNAm, either across or within traits (Methods). These results suggest that genetically-regulated DNAm loci that play a role in trait etiology correspond to lowly-methylated CpG sites in active chromatin regions, and that both gains and losses of DNAm can be pathogenic depending on the context.

Typically, variants contributing to the genetic basis of a trait are thought to act by affecting gene regulation. Concordantly, we observe an enrichment (Fisher's exact test P < 0.05) of trait-linked mCpGs in gene regulatory elements (OR = 1.63, P < 2.2e-16), compared to mCpGs without an identified trait link; 71% (2,390/3,381) of trait-linked mCpGs fall into this category. However, mCpGs corresponding to mQTL-specific colocalizations show a divergent profile compared to e/mQTL-shared colocalizations (Fig. 6b). While e/mQTL-shared GWAS-colocalized mCpGs are depleted in distal enhancers (OR = 0.68) and enriched in gene body element regions (OR = 1.54 to 2.22), promoters (OR = 2.19) and proximal enhancers (OR = 2.12), mQTL-specific GWAS-colocalized mCpGs are enriched in both proximal (OR = 1.36) and distal (OR = 1.39) enhancers, but not in promoters. Our results suggest that distal regulatory elements play an important role in DNAm impact on genotype-phenotype associations.

Integration of trait-linked methylated loci with functional maps enables the identification of trait-associated candidate genes

To identify genes that could mediate the effect of mQTLs on human traits, we integrated mQTL-specific trait-linked mCpGs with curated promoter- and enhancer-gene target predictions [38,39] and eQTM associations generated herein (Methods). We identified 68% (1,307/1,911) of mCpGs as functionally linked to gene(s) (Supplementary Table 6). Among highly supported (by ≥ 3 mCpGs) cases, we identify both well-known gene-trait associations, such as APOB with cholesterol, TERT with cancer, and ABO with blood traits. We also identify lesser characterized trait-linked genes, such as RUNX1 with asthma and TMEM72 with red blood cell counts, with biological evidence compatible with predicted trait links (Supplementary Note). Our results suggest that integrating the mQTL-trait maps generated herein with additional functional genomic maps can enable the identification of trait-linked candidate genes.

Discussion

To our knowledge, our study is the most comprehensive work that has assessed and compared the genome-wide DNAm profile of highly-diverse, healthy human tissue types in a systematic manner. Moreover, most studies that have characterized DNAm in relation to gene expression patterns have focused on blood cells [22,28], with some exceptions [19]. Our eQTM catalog facilitates improved CpG annotation, and allows for the characterization of disease-altered CpG-gene regulation landscapes by contrasting diseased with non-diseased tissue profiled herein. Additionally, integration of multi-tissue eQTM maps with single-cell expression data allows for the prediction of cellular abundances, as recently shown[40].

This work substantially contributes to an enhanced molecular characterization of complex traits. Across most traits, the aggregated contribution of mQTLs to trait links is larger than eQTLs, despite mQTLs being derived from substantially smaller sample sets. Hence, we conclude that generally, trait-associated variants are more likely to result in detectable changes to DNA methylation than gene expression. Consequently, we demonstrate that mQTLs can untap a substantial amount of molecular links to traits otherwise missed by eQTL-GWAS colocalization approaches. For these cases, mQTLs provide evidence of regulatory mechanisms underlying GWAS findings in absence of eQTL-based links to specific genes, and may pinpoint putative candidate genes. That is, integration of trait-linked mQTL associations with functional genomic maps like curated enhancer-gene target predictions [38,39] and eQTM associations can guide the design of variant-to-function studies for complex traits.

The enrichment and large overlap (71%) of trait-linked mCpGs with regulatory elements suggests that while methylation may often play a role along the causal pathway to phenotype, its role on genotype-phenotype associations may generally involve a gene expression mechanism. However, the lack of observed eQTLs for mQTL-specific GWAS colocalizations raises questions about how genetically-regulated DNAm impacts trait-related biology. Unidentified eQTLs derived from unprofiled transcripts herein, e.g. microRNAs or genes with very low expression levels (undetectable by RNAseq), are expected to explain part of the missing links. However, we hypothesize that some of the observed mQTL-specific colocalizations include loci where DNAm is genetically co-regulated with gene expression only in a particular context - such as early development, stem cells, cancer - which causally impacts the trait; but only methylation - not gene expression - is identifiable beyond the causal context, i.e. in the differentiated cells and/or healthy tissues from GTEx. In that sense, we suggest that DNAm may be more stable across different contexts as compared to expression, and that regulatory QTL mapping during context-specific, dynamic cellular processes can reveal otherwise hidden regulatory variation that may be particularly relevant in disease [41]. Alternatively, DNAm-phenotype links may involve additional molecular phenotypes other than gene expression. Given that intragenic DNAm can influence splicing [42], it is possible that sQTLs contribute to identified mQTL-trait links. Taken together our results emphasize, supported by the trait-linked regulatory pleiotropy patterns observed, the importance of integrating multiple -omics to exhaustively pinpoint molecular links and candidate genes to traits.

The dataset generated herein not only constitutes the largest cohort with multi-tissue DNAm data generated to date, but allows for integrative -omics analyses by enhancing existing transcriptome- and proteome-based GTEx datasets [4,11,43], and provides the research community with a valuable resource to investigate the inherited susceptibility to human disease and complex traits from both cross-tissue and cross-omics perspectives. Altogether, our results contribute to a better understanding of the human DNA methylome, and its relationship with both the transcriptome and complex traits.

Methods

Obtention and processing of functional genomic data

Gene regulatory element annotations were derived from ENCODE Encyclopedia version 5 (ENCODE5) predicted cis-Regulatory Elements (cCREs) catalog, including distal enhancers [ENCFF535MKS], proximal enhancers [ENCFF036NSJ], promoter-like regions [ENCFF379UDA] and putative insulators [ENCFF262LCI], where [${id}] corresponds to ENCODE5 file id from https://www.encodeproject.org/ . Putative insulators are defined herein by CTCF binding sites [44] unrelated to enhancers, promoter-like regions and DNase-H3K4me3 marks. Gene body annotations were obtained from GENCODE version 26 website [https://www.gencodegenes.org/human/release_26.html]. Genomic variant annotations were derived from Ensembl build 102 VEP cache [ftp://ftp.ensembl.org/pub/release-102/variation/indexed_vep_cache/homo_sapiens_vep_102_GRCh38.tar.gz]. Different annotations were collapsed: splice region, acceptor and donor sites were collapsed to ‘splice site’ and coding sequence sites to ‘CDS’. Open chromatin annotations derived from DNase-seq were obtained from ENCODE project version 5 website [https://www.encodeproject.org/]. DNase-seq profiles of adult individuals matching tissues analyzed herein were selected, including breast [ENCFF788BHK], colon transverse [ENCFF903WEH], kidney [ENCFF407WZV], lung [ENCFF886KAA], muscle [ENCFF983ONG], ovary [ENCFF500HAK], prostate [ENCFF557QYU] and testis [ENCFF761JZU]. Chromatin state predictions corresponding to a 18-state model derived from 6 marks - H3K4me3, H3K4me1, H3K36me3, H3K27me3, H3K9me3 and H3K27ac - were obtained from ROADMAP FTP site [https://egg2.wustl.edu/roadmap/data/byFileType/chromhmmSegmentations/ChmmModels/core_K27ac/jointModel/final/]. Chromatin state predictions corresponding to adult individuals’ epigenomes matching tissues analyzed herein were selected, including colonic mucosa [E075], lung [E096], muscle skeletal [E108], ovary [E097] and primary mononuclear cells from peripheral blood [E062]. Transcription factor (TF) binding annotations were derived from ENCODE version 2 and 3 ChIP-seq clustered peaks combined from 1,256 experiments, representing 340 transcription factors in 129 cell and tissue types; obtained from UCSC table browser [http://genome.ucsc.edu/cgi-bin/hgTables , table encRegTfbsClustered, build hg38]. CpG island (CGI) predictions were derived from UCSC table browser [http://genome.ucsc.edu/cgi-bin/hgTables , table cpgIslandExt, build hg38]. Enhancer-gene predictions were derived from Genehancer predictions from UCSC table browser [http://genome.ucsc.edu/cgi-bin/hgTables , table geneHancerRegElementsDoubleElite, build hg38], and from https://www.engreitzlab.org/resources/ (all element-gene connections with ABC scores > = 0.015).

Generation of epigenome-wide DNAm methylation (DNAm) data

Epigenome-wide DNAm was derived from 1,000 tissue samples from 9 unique tissue types (Supplementary Table 1) obtained from 424 GTEx subjects. DNA samples were extracted from GTEx tissue samples using Qiagen Gentra Puregene method at GTEx Laboratory Data, Analysis and Coordinating Center (LDACC), and sent to the Institute for Population and Precision Health Laboratory on 96-well plates. Tissue types were not batched by plate. Bisulfite conversion was applied to 500 ng of DNA using EZ-96 DNA methylation kit (Zymo Research, Irvine, CA, USA). All samples were then prepared and analyzed in accordance with the manufacturer guidelines and protocol for the Infinium MethylationEPIC array (Illumina, San Diego, CA, USA); which was utilized to measure DNAm levels for 866,895 (867K) genomic locations, encompassing primarily CpG dinucleotides but also non-CpG sites. For simplicity, we refer to both types of sites as CpGs or CpG sites. Raw DNAm data was processed with ChAMP software [45]. For sample quality control (QC), we excluded (1) 3 samples with undetectable or missing methylation values (detection $P > 0.01$) in ≥ 5% of CpGs and (2) 6 samples with mismatched sex. For CpG QC, we excluded (1) CpGs that had detection P > 0.01 in ≥ 1 samples (N = 44,135) and that had a beadcount < 3 in ≥ 5% samples (N = 660), (2) cross-reactive CpGs (N = 40,812), (3) variant-overlapping CpGs or within a single base pair extension (N = 7,708) [46] and (4) CpGs mapping to sex chromosomes. A total of 754,054 CpGs passed QC, could be mapped and lifted over from GRCh37 to GRCh38 human genome build and were retained for further analysis.

Raw DNAm values were background-adjusted using the single sample normal-exponential out-of-band (*ssnoob*) method with dye bias correction [47,48]. DNAm β values were normalized using the beta mixture quantile (*BMIQ*) method, adjusting for type I/II probe bias [49]; output β values for each CpG were used in downstream analysis. After normalization, we removed one additional sample with array-derived genotype profile not matching WGS-derived one. Principal component analysis (PCA) was conducted on DNAm β values within each tissue type, and 3 samples were removed for being outliers with respect to the top 5 principal components (PCs) of the corresponding tissue. A total of 987 samples passed final QC and were retained for further analysis.

Estimation of tissue similarity based on DNAm and gene expression

Mean DNAm and gene expression levels for each tissue type were used to calculate tissue distances (1 - Spearman rank-correlation coefficient ( )) in a pairwise fashion. For DNAm, was calculated on the mean DNAm for each CpG (across all samples for a given tissue) in M-value units; and each CpG was weighted by its cross-tissue variance. The entire set of post-QC profiled CpGs (N = 754,054) was considered for the analysis. For gene expression, was calculated on the mean expression for each gene (across all samples for a given tissue) in log2(x + 1) transformed transcript per million units (TPM); and each gene was weighted by its cross-tissue variance. Only genes expressed in at least one of the nine tissues (N = 37,686) were considered for the analysis. We performed hierarchical clustering of the tissues using pvclust 2.0.0 [50] with complete linkage and nboot = 1,000.

Expression quantitative trait methylation (eQTM) mapping

We define an eQTM as the association of CpG DNAm with proximal gene expression, considering a ± 1.5 Mb window centered on the CpG locus, thereby enabling the identification of short- and long-range CpG-gene associations. We analyzed samples with available DNAm, expression and genotype data (Supplementary Table 1), comprising a total of 490 samples, from 25 - testis - to 131 - lung - per tissue, excluding kidney cortex due to insufficient sample size (N = 5). We obtained DNAm residuals by regressing DNAm-derived PEER factors [51] used in *cis*-mQTL mapping (see below) from inverse-normalized-transformed DNAm levels, and gene expression residuals by regressing expression-derived PEER factors used for eQTL mapping in [4] from inverse-normalized-transformed gene expression levels. For each CpG site in each tissue, we calculated Spearman correlation of DNAm with gene expression residuals of proximal (± 1.5 Mb window) genes. We applied Bonferroni multiple testing correction to nominal p-values, accounting for multiple genes tested per CpG. Then, we adjusted for multiple CpGs tested by applying q-value multiple testing correction [52] to the set of top Bonferroni-adjusted p-values per CpG. We defined significant CpGs involved in eQTMs (eCpGs) at FDR < 0.05, and for those, significant eQTMs were defined at Bonferroni-adjusted p-value < 0.05. To overcome QTM-mapping limited power due to per-tissue available sample sizes, we used an approach to perform a cross-tissue QTL analysis by leveraging QTL signal across tissues [29], implemented in the R package *mashr*. We assessed eQTM replication for all tissues in the FUSION Skeletal Muscle Study cohort [19]. Replication was assessed by means of π1, which enables the estimation of the true positive rate of findings derived from a discovery dataset in a replication dataset [53]. See additional details of eQTM identification and characterization in Supplementary Note.

Quantitative trait loci (QTL) mapping for methylation and gene expression QTLs (mQTLs, eQTLs)

We define mQTLs as proximal variants, i.e. in *cis*, to a CpG with a significant genotype effect on its DNAm estimates, considering a ± 500 Kb window from the CpG locus. To assess mQTLs, we considered QC-ed inverse-normalized DNAm data, generated and presented here as part of the eGTEx project, and QC-ed genotype data derived from GTEx v8 [4] filtered at variant minor allele frequency (MAF) > 0.01 per tissue. For each variant-CpG pair, we fit a linear regression model separately in each tissue, and tested for significance of genotype on methylation estimates while adjusting for additional known and unknown factors (see Supplementary Note). We implemented the model in an adaptation of *FastQTL* [54], available at https://github.com/broadinstitute/gtex-pipeline/tree/master/qtl, and corrected for multiple testing of variants per CpG [54] and multiple CpGs tested [53], defining significant mQTL CpGs (mCpGs) at FDR < 0.05. An equivalent approach was utilized to identify eQTLs. Conditional QTL analysis was employed to identify multiple independent mQTLs and eQTLs for mCpGs and eGenes, respectively; as well as corresponding lead variants for each independent QTL locus. This approach accounts for allelic heterogeneity (AH), where distinct genetic variants at a locus simultaneously and independently affect methylation at a given CpG site. For that, we applied a stepwise regression procedure as described in [30]; see details in Supplementary Note. Similarly as in eQTM mapping, we utilized *mashr* to identify mQTLs and eQTLs detectable after leveraging QTL signal across tissues. Across the article, we refer to CpGs and genes with at least one significant mQTL or eQTL as mCpGs and eGenes, respectively; and to mQTL and eQTL significant variants as mVariants and eVariants, respectively. We assessed mQTL replication in the BEST blood cohort [26] and the FUSION Skeletal Muscle Study cohort [19].

See additional details of mQTL and e QTL identification and characterization in Supplementary Note.

QTL enrichment in genomic annotations

Functional enrichment analyses were performed using *torus* [55] , similarly to [4] . In brief, the command “torus -d ${qtl_statistics} -annot ${annotation_file} -est –fastqtl” was utilized; where ${qtl_statistics} correspond to QTL-mapping data for eQTLs (full eQTL-tested gene set per tissue) or mQTLs (subset of 21k mQTL-tested CpGs), and ${annotation_file} corresponds to QTL-tested annotated variants. Variants were annotated with Ensembl’s Variant Effect Predictor (VEP) utilizing datasets from several sources: gene regulatory element and open chromatin annotations were obtained from ENCODE5, gene body annotations were obtained from Ensembl, chromatin state predictions from ROADMAP, TF binding and CGI annotations were obtained from from UCSC browser (see Obtention and processing of functional genomic data). Enrichment across tissues was evaluated by modeling single-tissue enrichment estimates (log of odds ratio) with a random-effects model (*rma* function, *metafor* R package). Single-tissue and cross-tissue enrichment estimates are provided in Supplementary Table 3.

Colocalization of mQTLs with eQTLs

We investigated the associations between mQTLs and eQTLs by means of QTL effect size colocalization with *coloc* [56] using default priors. For each significantly (PP4 > 0.5) colocalized mQTL-eQTL pair in each tissue, the top-colocalized e/mVariant was defined as the one with the largest PP4 value. See additional details of mQTL-eQTL colocalization in Supplementary Note.

Characterization of colocalized mQTL-eQTL loci

mQTL-eQTL concordance in direction of effects

A mQTL-eQTL pair was defined as concordant if the mQTL sign of the top-colocalized e/mVariant matched the corresponding eQTL sign and discordant otherwise. An exact binomial test was conducted to assess whether the proportion of discordant/concordant cases differed significantly from 50%, and the null hypothesis (proportion = 53%, P < 2.2e-16) was rejected. Subsequently, we assessed whether the discordance/concordance rate varied as a function of mCpG location in gene regulatory regions, considering promoters and proximal enhancers jointly, distal enhancers and insulators. Gene regulatory element annotations were derived from ENCODE5 cCREs catalog (see Obtention and processing of functional genomic data) and details of mCpG annotation are provided in Supplementary Note.

Identification of mQTL-eQTL regulatory pleiotropy

Considering QTLs, we define regulatory pleiotropy as the event of a variant or a set of variants in a QTL region impacting multiple eGenes and/or mCpGs. In order to characterize the pleiotropic nature of mQTLs, we quantified the number of mCpGs and eGenes involved in loci harbouring eQTL-mQTL (e/mQTL) colocalizations, and classified mCpGs involved in at least one significant e/mQTL colocalization (PP4 > 0.50) in four tiers, depending on their mCpG and eGene connectivity level (Supplementary Fig. 4). Of note, no inference of the nature of the pleiotropic effect, whether vertical or horizontal, is made in this classification. Tiers are defined as follows; a) Tier 1: mCpGs that colocalize with a single eGene and vice versa (1:1 connectivity), b) Tier 2: mCpGs that colocalize with multiple eGenes, and each one of those eGenes uniquely colocalize with that single mCpG (1:m connectivity), c) Tier 3: mCpGs that colocalize with a single eGene, which colocalizes with multiple mCpGs (n:1 connectivity) and d) Tier 4: mCpGs that colocalize with a multiple eGenes, where at least one of the eGenes colocalize with multiple mCpGs (n:m connectivity).

Colocalization of QTL with GWAS signal

Colocalization of ovary cancer GWAS with QTL signal of HOXD pleiotropic locus

We hypothesized that the mCpGs and eGenes in the HOXD region may be linked to ovarian cancer risk, and tested it by means of QTL-GWAS colocalization. Ovary cancer GWAS summary statistics [57] were obtained from the Ovarian Cancer Association Consortium (OCAC) website http://ocac.ccge.medschl.cam.ac.uk/ and were filtered - not imputed - as in [5]. Considering OCAC GWAS along with QTL statistics from the set of pleiotropic mCpGs and eGenes identified in the HOXD locus (Fig. 3a), we performed colocalization analysis with *coloc* [56] using default priors. Considering QTLs statistics, colocalization was performed based on effect size and associated standard error values; p-values and corresponding variant MAFs were used for OCAC GWAS data. The probability of one causal variant associated with both traits (PP4) was used to identify significant (PP4 > 0.50) colocalizations.

Determination of GWAS significant loci

To investigate possible associations between genetically regulated molecular and complex traits, including disease and ‘healthy’ phenotypes, we employed GWAS summary statistics of 87 GWASs; data production and quality control are described in detail in [5]. We identify significant GWAS hit loci (i.e. genomic windows containing GWAS signal) similarly as described in [5]. In brief, the GWAS summary statistics were split into 1,702 approximately LD-independent regions [58] (Supplementary Table 5). Each region was categorized as a significant GWAS hit locus (GWAS hit) provided it encompassed a non-imputed GWAS significant ($P < 5 \times 10^{-8}$) variant.

Determination of QTL-GWAS significant loci

For each GWAS and each QTL tissue pair, colocalization was performed with *coloc* and *fastenloc*. The latter is an improved version of *enloc* [59] and was recently described in multiple works [60,61].

coloc approach

For each GWAS, at each GWAS locus, we identified overlapping (> 1 bp) mCpG loci from each of the 9 analyzed tissues, considering per-tissue significant (FDR < 0.05) mCpGs resulting from the single-tissue QTL-mapping approach. For each overlapping mCpG-GWAS region pair, we applied *coloc* [56] to mQTL along with GWAS summary statistics. For mCpGs with secondary QTLs, i.e. multiple independent mQTLs, conditional - to independent lead mVariants - QTL signals were also tested for colocalization. By this, we prevent putative colocalizations to be missed or miscalculated, since *coloc* assumes a single variant to be causal of QTL-GWAS effects [61]. Prior probabilities of a variant yielding a) a mQTL association (p1), b) a GWAS association (p2) and b) a mQTL and a GWAS association (p12) were estimated from *fastenloc* enrichment values in an analogous manner as done with *enloc* in [5] and provided in Supplementary Table 5. Only the regions with at least 50 variants in common between the GWAS and mCpG loci were tested for colocalization. Both for QTLs and GWAS statistics, colocalization was performed on effect size (*effect size*) and associated standard error (*effect size s.e.*) values. Used GWAS statistics were imputed from available z-score (*z*), allele frequency (*f*) and sample size values (*N*) by *effect size* ≈ *z*/(*f*(1 − *f*)*N*)^(1/2) and *effect size s.e.* ≈ *effect size* / *z*.

fastenloc approach

First, GWAS posterior inclusion probability (PIP) values were obtained with *torus* from available z-scores. Then, for each tissue, the DNAm levels, genotypes and mQTL-mapping covariates for CpGs and corresponding *cis* windows considered in the mQTL analysis were processed with *dap-g* [62]. Next, we used *fastenloc* to obtain regional colocalization probabilities (RCP) for all tuples of interest (GWAS hit, trait, tissue, mCpG) by subsetting corresponding GWAS hits, and significant (single-tissue mQTL set: FDR < 0.05) mCpGs from the genome-wide *fastenloc* output. An equivalent approach was employed for eQTL-GWAS colocalization; the *dap-g* pre-computed eQTL annotations from GTEx (v8) data were obtained from the fastenloc github repository: https://github.com/xqwen/fastenloc. Evaluation of mQTL-GWAS colocalization approach is described in Supplementary Note.

Characterization of signatures of trait-linked mCpGs

Tissue-specific enrichment in mQTL-GWAS colocalizations

To identify traits with a tissue-specific colocalization enrichment profile, we performed a multi-sample test for equality of proportions without continuity correction. For each trait with > 5 mQTL-GWAS colocalizations, we compared the observed proportions of mQTL-GWAS colocalizations per tissue to the overall proportion of colocalizations for all tissues. We identified traits with a disproportionate amount of colocalizations in at least one tissue at Bonferroni-adjusted P < 0.05. For these traits, scaled-by-trait colocalization proportions per tissue are shown in Fig. 6a, where tissues and traits were clustered via complete-linkage hierarchical clustering based on euclidean distance. A two-sided Fisher's exact test was conducted to determine significant tissue-trait enrichments at Bonferroni-adjusted P < 0.01, which are indicated by crossed cells in Fig. 6a. This approach has several limitations. We did not observe tissues of relevance for several traits, possibly due to the examination of mQTLs derived from only 9 tissues. For certain tissue-trait pairs, the limitedness of the mQTL catalog examined herein, and the low number of colocalizations identified, may have also impacted the completeness and accuracy of the identified tissue-specific enrichment profiles. This analysis would benefit from a more powered and exhaustive mQTL catalog.

Characterization of molecular regulatory pleiotropy of trait-linked mCpGs

We evaluated the enrichment of mVariants corresponding to mQTL-GWAS colocalizations in mVariants regulating multiple versus single mCpGs. For each mCpG tested for mQTL-GWAS colocalization, we kept the corresponding colocalization result with the highest *coloc* PP4 value and the corresponding mVariant. We classified mVariants as pleiotropic if they were estimated to be causal of multiple mCpGs, or as non-pleiotropic, if they were causal of a single mCpG. Causality was defined considering *dap-g* fine-mapping estimates utilized in the *fastenloc* mQTL-GWAS colocalization approach, and mQTL credible sets were defined at 90% confidence. We classified mVariants as being also eVariants if they were estimated to be causal of at least one eGene, estimating eQTL credible sets in an analogous manner to mQTL sets. Significant enrichment of trait-linked versus trait-unlinked mVariants in multiple-mCpG and eVariant sets was defined at two-sided Fisher's exact test P < 0.05. Additionally, we subsetted mCpGs tested for mQTL-GWAS colocalization that were involved in e/mQTL colocalizations, and classified them into mCpG-eGene pleiotropy tiers (Supplementary Fig. 4). Enrichment of trait-linked versus trait-unlinked eGene and mCpGs in each tier was estimated at two-sided Fisher's exact test P < 0.05.

Characterization of DNAm signatures of trait-linked mCpGs

To evaluate the overlap of trait-linked mCpGs with open chromatin regions, we utilized annotations derived from ENCODE5 open chromatin regions (see Obtention and processing of functional genomic data). See additional details in Supplementary Note.

We evaluated the putative enrichment of increased disease-risk alleles in gains or losses of DNAm, both across and within GWAS traits. From the set of 87 GWAS tested for mQTL colocalization, we subsetted colocalization results corresponding to 31 GWAS disease traits with for which at least one significant mQTL colocalization was identified. For each GWAS trait and mCpG tested for mQTL-GWAS colocalization, we kept the corresponding colocalization result with the highest *coloc* PP4 value. For each GWAS trait, we classified tested mCpGs into methylation gain or loss categories, based on the sign of the allelic mQTL effect associated with increased disease risk, i.e. positive GWAS effect. Enrichment of each disease in methylation gains and losses was determined at Bonferroni-corrected two-sided Fisher's exact test $P < 0.05$. No disease was identified as enriched for DNAm gains or losses. Despite individual traits not being significantly enriched for DNAm change biases, it is possible that a global bias could be detected when considering all traits at once. To test this hypothesis, each disease was labelled as 'DNAm gain' if the corresponding number of colocalized mCpGs associated to a DNAm gain outnumbered the DNAm-loss ones, and labelled as 'DNAm-loss' otherwise. An exact binomial test was performed to assess whether the proportion of DNAm-gained or -loss labelled diseases deviated significantly ($P < 0.05$) from 50%, which was not the case ($P = 0.86$).

Characterization of trait-linked mCpGs overlap with gene regulatory regions

Gene regulatory element annotations were derived from ENCODE5 cCREs catalog and gene body element annotations were derived from obtained from GENCODE (see Obtention and processing of functional genomic data). To annotate mCpGs for cCRE and gene body elements, we extended the span of their genomic location by +/- 100bps, and checked for overlap (> = 1bp) with element regions. Trait-linked mCpGs were classified as eQTL-shared or mQTL-specific (see Supplementary Note: Evaluation of mQTL-GWAS colocalization). Enrichment significance of eQTL-shared or mQTL-specific trait-linked mCpGs in each gene regulatory-region or gene body element category was estimated at Fisher's exact test $P < 0.05$.

Characterization of trait-linked mCpGs overlap with functional maps

To evaluate the overlap of mQTL-specific trait-linked mCpGs with genomic regions linked to specific genes, we considered functional maps based on curated regulatory-region to gene target predictions [38,39] and eQTM predictions generated herein. We extended the genomic location span of mCpGs by +/- 100bps, and checked for overlap (> = 1bp) with eCpGs and regulatory regions. For each mCpG, we annotated corresponding regulatory-region gene targets and eCpG-correlated genes; a particular gene was annotated to a mCpG if overlap was found in one or more functional maps. Annotations were collapsed at a GWAS hit level by adding up the number of predicted mCpG-gene links corresponding to the locus, hence providing an estimate of mCpGs that support a gene candidate per trait-associated locus (Supplementary Table 6).

Data availability

All GTEx open-access data, including summary statistics of mQTLs, will be available soon on the GTEx Portal (https://gtexportal.org/home/datasets). All GTEx protected data are available via dbGaP (accession phs000424.v8). Access to the raw sequence data is provided through the AnVIL platform (https://gtexportal.org/home/protectedDataAccess). Full summary statistics of mQTLs, eQTMs and QTL-GWAS colocalizations will be available upon acceptance (zenodo repository).

Code availability

All analyses were performed with existing open access methods. The adaptation of *FastQTL* is available at https://github.com/broadinstitute/gtex-pipeline/tree/master/qtl. *fastenloc* is available at https://github.com/xqwen/fastenloc (version 1.0 was used). Scripts for conditional QTL mapping are available at https://github.com/funpopgen/multiple_eqtl_mapping. *torus* is available at https://github.com/xqwen/torus (version 1.0.0.dev was used). mashr is available at https://github.com/stephenslab/mashr (version 0.2.6 was used).

Declarations

Data availability

All GTEx open-access data, including summary statistics of mQTLs, will be available soon on the GTEx Portal (https://gtexportal.org/home/datasets). All GTEx protected data are available via dbGaP (accession phs000424.v8). Access to the raw sequence data is provided through the AnVIL platform (https://gtexportal.org/home/protectedDataAccess). Full summary statistics of mQTLs, eQTMs and QTL-GWAS colocalizations will be available upon acceptance (zenodo repository).

Code availability

All analyses were performed with existing open access methods. The adaptation of *FastQTL* is available at https://github.com/broadinstitute/gtex-pipeline/tree/master/qtl. *fastenloc* is available at https://github.com/xqwen/fastenloc (version 1.0 was used). Scripts for conditional QTL mapping are available at https://github.com/funpopgen/multiple_eqtl_mapping. *torus* is available at https://github.com/xqwen/torus (version 1.0.0.dev was used). mashr is available at https://github.com/stephenslab/mashr (version 0.2.6 was used).

Acknowledgments

We thank the donors and their families for their generous gifts of biospecimens to the GTEx research project; the Genomics Platform at the Broad Institute for data generation; D. Delgado and L. Tong for assistance with assessing mQTL replication, and M. Goraj and J. Witkos for comments on an earlier version of the manuscript. This work was completed in part with computational resources provided by the Center for Research Informatics at the University of Chicago and the Centre for Genomic Regulation.

Funding

Supported by U01 HG007601 (to B.L.P.) and R35ES028379 (to B.L.P.); and 2R01 GM108711 (to L.S.C) and U24 CA210993-SUB (to L.S.C).

Author contributions

B.P. conceived the study; M.O. conceived and led all analysis supervised by B.L.P. and L.S.C.; M.O. performed all bioinformatic analysis, granted K.D. and Y.L. contributions; M.O. led the writing and editing of the manuscript and supplement; B.L.P., L.S.C. and H.A. contributed to the editing of the manuscript and supplement; M.O., B.L.P. and L.S.C. coordinated analyses of all contributing authors; F.J. generated the DNA methylation data; M.G.K. supervised the generation of the DNA methylation data; K.D. processed and QC-ed the DNA methylation data; Y.L. contributed to the mQTL functional characterization analysis. All authors read and approved the final manuscript.

Competing interests

None to declare.

References

1. Nicolae, D. L. *et al.* Trait-associated SNPs are more likely to be eQTLs: annotation to enhance discovery from GWAS. *PLoS Genet.* **6**, e1000888 (2010).
2. Cano-Gamez, E. & Trynka, G. From GWAS to Function: Using Functional Genomics to Identify the Mechanisms Underlying Complex Diseases. *Front. Genet.* **11**, 424 (2020).
3. GTEx Consortium *et al.* Genetic effects on gene expression across human tissues. *Nature* **550**, 204–213 (2017).
4. GTEx Consortium. The GTEx Consortium atlas of genetic regulatory effects across human tissues. *Science* **369**, 1318–1330 (2020).
5. Barbeira, A. N. *et al.* Exploiting the GTEx resources to decipher the mechanisms at GWAS loci. *Genome Biol.* **22**, 49 (2021).
6. Ongen, H. *et al.* Estimating the causal tissues for complex traits and diseases. *Nat. Genet.* **49**, 1676–1683 (2017).
7. Gamazon, E. R. *et al.* Using an atlas of gene regulation across 44 human tissues to inform complex disease- and trait-associated variation. *Nat. Genet.* **50**, 956–967 (2018).
8. Banovich, N. E. *et al.* Methylation QTLs are associated with coordinated changes in transcription factor binding, histone modifications, and gene expression levels. *PLoS Genet.* **10**, e1004663 (2014).
9. Melé, M. *et al.* Human genomics. The human transcriptome across tissues and individuals. *Science* **348**, 660–665 (2015).
10. Cancer Genome Atlas Research Network *et al.* The Cancer Genome Atlas Pan-Cancer analysis project. *Nat. Genet.* **45**, 1113–1120 (2013).
11. eGTEx Project. Enhancing GTEx by bridging the gaps between genotype, gene expression, and disease. *Nat. Genet.* **49**, 1664–1670 (2017).
12. Li, E., Beard, C. & Jaenisch, R. Role for DNA methylation in genomic imprinting. *Nature* **366**, 362–365 (1993).
13. Payer, B. & Lee, J. T. X chromosome dosage compensation: how mammals keep the balance. *Annu. Rev. Genet.* **42**, 733–772 (2008).
14. Maurano, M. T. *et al.* Role of DNA Methylation in Modulating Transcription Factor Occupancy. *Cell Rep.* **12**, 1184–1195 (2015).
15. Jin, Z. & Liu, Y. DNA methylation in human diseases. *Genes Dis* **5**, 1–8 (2018).
16. Kaminsky, Z. A. *et al.* DNA methylation profiles in monozygotic and dizygotic twins. *Nat. Genet.* **41**, 240–245 (2009).
17. Hannon, E. *et al.* Methylation QTLs in the developing brain and their enrichment in schizophrenia risk loci. *Nat. Neurosci.* **19**, 48–54 (2016).
18. Morrow, J. D. *et al.* Human Lung DNA Methylation Quantitative Trait Loci Colocalize with Chronic Obstructive Pulmonary Disease Genome-Wide Association Loci. *Am. J. Respir. Crit. Care Med.* **197**, 1275–1284 (2018).
19. Taylor, D. L. *et al.* Integrative analysis of gene expression, DNA methylation, physiological traits, and genetic variation in human skeletal muscle. *Proc. Natl. Acad. Sci. U. S. A.* **116**, 10883–10888 (2019).
20. Huan, T. *et al.* Genome-wide identification of DNA methylation QTLs in whole blood highlights pathways for cardiovascular disease. *Nat. Commun.* **10**, 4267 (2019).
21. Gibbs, J. R. *et al.* Abundant quantitative trait loci exist for DNA methylation and gene expression in human brain. *PLoS Genet.* **6**, e1000952 (2010).
22. Gutierrez-Arcelus, M. *et al.* Passive and active DNA methylation and the interplay with genetic variation in gene regulation. *Elife* **2**, e00523 (2013).
23. Gutierrez-Arcelus, M. *et al.* Tissue-specific effects of genetic and epigenetic variation on gene regulation and splicing. *PLoS Genet.* **11**, e1004958 (2015).
24. Bell, J. T. *et al.* DNA methylation patterns associate with genetic and gene expression variation in HapMap cell lines. *Genome Biol.* **12**, R10 (2011).

25. McClay, J. L. *et al.* High density methylation QTL analysis in human blood via next-generation sequencing of the methylated genomic DNA fraction. *Genome Biol.* **16**, 291 (2015).

26. Pierce, B. L. *et al.* Co-occurring expression and methylation QTLs allow detection of common causal variants and shared biological mechanisms. *Nat. Commun.* **9**, 804 (2018).

27. Bommarito, P. A. & Fry, R. C. Chapter 2-1 - The Role of DNA Methylation in Gene Regulation. in *Toxicoepigenetics* (eds. McCullough, S. D. & Dolinoy, D. C.) 127–151 (Academic Press, 2019).

28. Kim, S. *et al.* Expression Quantitative Trait Methylation Analysis Reveals Methylomic Associations With Gene Expression in Childhood Asthma. *Chest* **158**, 1841–1856 (2020).

29. Urbut, S. M., Wang, G., Carbonetto, P. & Stephens, M. Flexible statistical methods for estimating and testing effects in genomic studies with multiple conditions. *Nat. Genet.* **51**, 187–195 (2019).

30. Brown, A. A. *et al.* Predicting causal variants affecting expression by using whole-genome sequencing and RNA-seq from multiple human tissues. *Nat. Genet.* **49**, 1747–1751 (2017).

31. Goode, E. L. *et al.* A genome-wide association study identifies susceptibility loci for ovarian cancer at 2q31 and 8q24. *Nat. Genet.* **42**, 874–879 (2010).

32. Kar, S. P. *et al.* Network-Based Integration of GWAS and Gene Expression Identifies a HOX-Centric Network Associated with Serous Ovarian Cancer Risk. *Cancer Epidemiol. Biomarkers Prev.* **24**, 1574–1584 (2015).

33. Shah, N. & Sukumar, S. The Hox genes and their roles in oncogenesis. *Nat. Rev. Cancer* **10**, 361–371 (2010).

34. Bojesen, S. E. *et al.* Multiple independent variants at the TERT locus are associated with telomere length and risks of breast and ovarian cancer. *Nat. Genet.* **45**, 371–84, 384e1–2 (2013).

35. Huan, T. *et al.* Genome-wide identification of microRNA expression quantitative trait loci. *Nat. Commun.* **6**, 6601 (2015).

36. Giambartolomei, C. *et al.* A Bayesian framework for multiple trait colocalization from summary association statistics. *Bioinformatics* **34**, 2538–2545 (2018).

37. Gleason, K. J., Yang, F., Pierce, B. L., He, X. & Chen, L. S. Primo: integration of multiple GWAS and omics QTL summary statistics for elucidation of molecular mechanisms of trait-associated SNPs and detection of pleiotropy in complex traits. *Genome Biol.* **21**, 236 (2020).

38. Nasser, J. *et al.* Genome-wide enhancer maps link risk variants to disease genes. *Nature* 1–6 (2021).

39. Fishilevich, S. *et al.* GeneHancer: genome-wide integration of enhancers and target genes in GeneCards. *Database* **2017**, (2017).

40. Teschendorff, A. E., Zhu, T., Breeze, C. E. & Beck, S. EPISCORE: cell type deconvolution of bulk tissue DNA methylomes from single-cell RNA-Seq data. *Genome Biol.* **21**, 221 (2020).

41. Umans, B. D., Battle, A. & Gilad, Y. Where Are the Disease-Associated eQTLs? *Trends Genet.* **37**, 109–124 (2021).

42. Maunakea, A. K., Chepelev, I., Cui, K. & Zhao, K. Intragenic DNA methylation modulates alternative splicing by recruiting MeCP2 to promote exon recognition. *Cell Res.* **23**, 1256–1269 (2013).

43. Jiang, L. *et al.* A Quantitative Proteome Map of the Human Body. *Cell* (2020) doi:10.1016/j.cell.2020.08.036.

44. Ali, T., Renkawitz, R. & Bartkuhn, M. Insulators and domains of gene expression. *Curr. Opin. Genet. Dev.* **37**, 17–26 (2016).

45. Morris, T. J. *et al.* ChAMP: 450k Chip Analysis Methylation Pipeline. *Bioinformatics* **30**, 428–430 (2014).

46. Pidsley, R. *et al.* Critical evaluation of the Illumina MethylationEPIC BeadChip microarray for whole-genome DNA methylation profiling. *Genome Biol.* **17**, 208 (2016).

47. Fortin, J.-P., Triche, T. J., Jr & Hansen, K. D. Preprocessing, normalization and integration of the Illumina HumanMethylationEPIC array with minfi. *Bioinformatics* **33**, 558–560 (2017).

48. Triche, T. J., Jr, Weisenberger, D. J., Van Den Berg, D., Laird, P. W. & Siegmund, K. D. Low-level processing of Illumina Infinium DNA Methylation BeadArrays. *Nucleic Acids Res.* **41**, e90 (2013).

49. Teschendorff, A. E. *et al.* A beta-mixture quantile normalization method for correcting probe design bias in Illumina Infinium 450 k DNA methylation data. *Bioinformatics* **29**, 189–196 (2013).

50. Suzuki, R. & Shimodaira, H. *pvclust: Hierarchical Clustering with P-Values via Multiscale Bootstrap Resampling*. (2015).

51. Stegle, O., Parts, L., Piipari, M., Winn, J. & Durbin, R. Using probabilistic estimation of expression residuals (PEER) to obtain increased power and interpretability of gene expression analyses. *Nat. Protoc.* **7**, 500–507 (2012).

52. Leek, J. T. & Storey, J. D. Capturing heterogeneity in gene expression studies by surrogate variable analysis. *PLoS Genet.* **3**, 1724–1735 (2007).

53. Storey, J. D. & Tibshirani, R. Statistical significance for genomewide studies. *Proc. Natl. Acad. Sci. U. S. A.* **100**, 9440–9445 (2003).

54. Ongen, H., Buil, A., Brown, A. A., Dermitzakis, E. T. & Delaneau, O. Fast and efficient QTL mapper for thousands of molecular phenotypes. *Bioinformatics* **32**, 1479–1485 (2016).

55. Wen, X. Molecular QTL discovery incorporating genomic annotations using Bayesian false discovery rate control. *Ann. Appl. Stat.* **10**, 1619–1638 (2016).

56. Giambartolomei, C. *et al.* Bayesian test for colocalisation between pairs of genetic association studies using summary statistics. *PLoS Genet.* **10**, e1004383 (2014).

57. Phelan, C. M. *et al.* Identification of 12 new susceptibility loci for different histotypes of epithelial ovarian cancer. *Nat. Genet.* **49**, 680–691 (2017).

58. Berisa, T. & Pickrell, J. K. Approximately independent linkage disequilibrium blocks in human populations. *Bioinformatics* **32**, 283–285 (2016).

59. Wen, X., Pique-Regi, R. & Luca, F. Integrating molecular QTL data into genome-wide genetic association analysis: Probabilistic assessment of enrichment and colocalization. *PLoS Genet.* **13**, e1006646 (2017).

60. Pividori, M. *et al.* PhenomeXcan: Mapping the genome to the phenome through the transcriptome. *bioRxiv* 833210 (2019) doi:10.1101/833210.

61. Hukku, A. *et al.* Probabilistic colocalization of genetic variants from complex and molecular traits: promise and limitations. *Am. J. Hum. Genet.* **108**, 25–35 (2021).

62. Wen, X., Lee, Y., Luca, F. & Pique-Regi, R. Efficient Integrative Multi-SNP Association Analysis via Deterministic Approximation of Posteriors. *Am. J. Hum. Genet.* **98**, 1114–1129 (2016).

63. van der Maaten, L. Accelerating t-SNE using Tree-Based Algorithms. *J. Mach. Learn. Res.* **15**, 3221–3245 (2014).

64. Yu, Y., Wang, L. & Gu, G. The correlation between Runx3 and bronchial asthma. *Clin. Chim. Acta* **487**, 75–79 (2018).

65. Lindgren, D. *et al.* Cell-Type-Specific Gene Programs of the Normal Human Nephron Define Kidney Cancer Subtypes. *Cell Rep.* **20**, 1476–1489 (2017).

66. Babitt, J. L. & Lin, H. Y. Mechanisms of anemia in CKD. *J. Am. Soc. Nephrol.* **23**, 1631–1634 (2012).

Figures

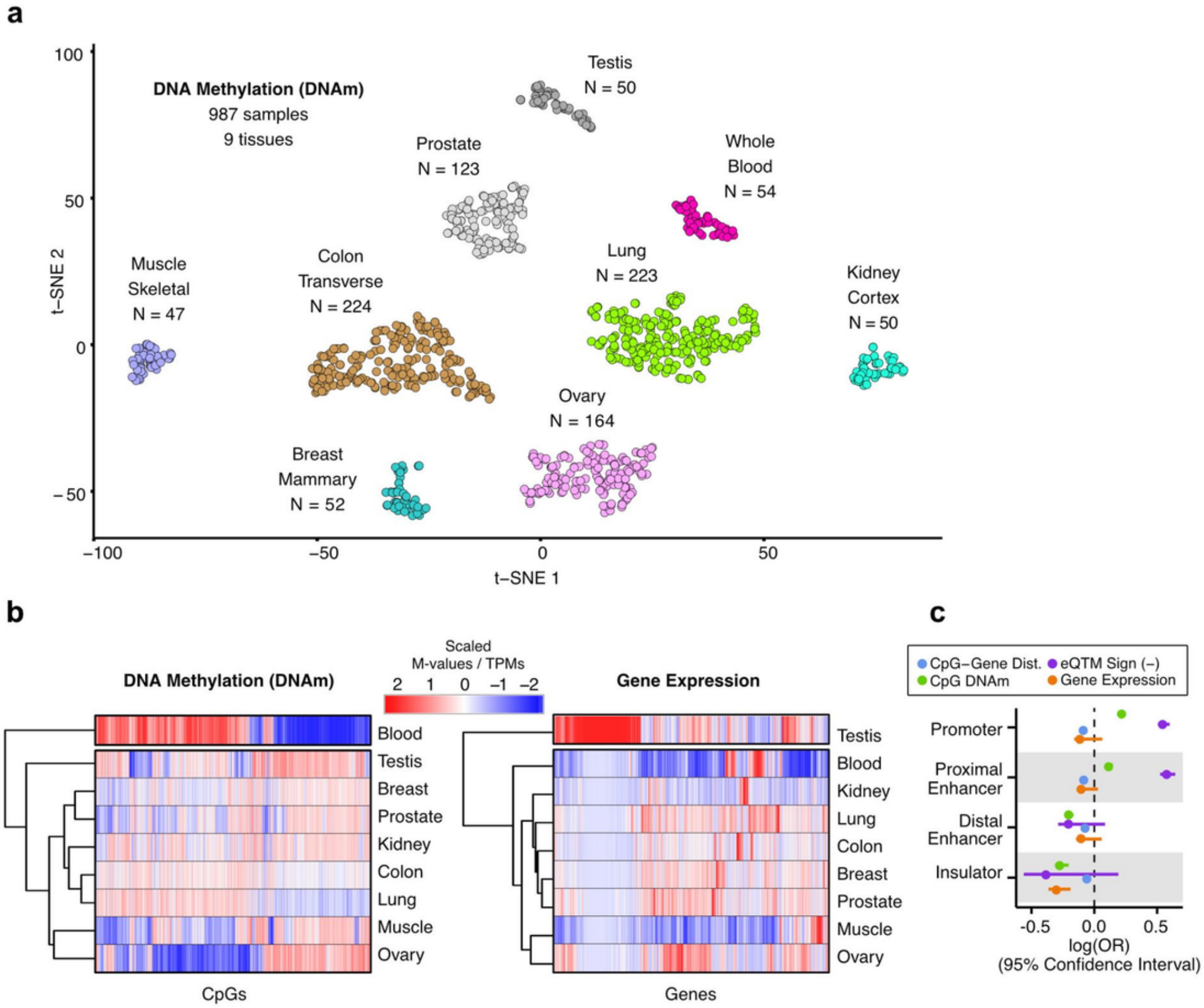

Figure 1

Characterization of methylomes across tissues. (a) Sample similarity based on DNAm profiles. Dimensionality reduction was performed with a t-Distributed Stochastic Neighbor Embedding approach (t-SNE). (b) Hierarchical tissue clustering based on complete methylomes (left panel) and transcriptomes (right panel) of nine tissues (x axis). The molecular phenotypes displayed (y axis) correspond to the top 20,000 most divergent CpG sites and genes across tissues. DNAm and gene expression values are column-wise scaled. (c) Contribution (x axis, square-root transformed log(OR)) of selected factors to eQTM likelihood (presence) for different gene regulatory elements (y axis). Dist.: Distance. OR: Odds Ratio. Factor units: CpG-gene distance [Kb], eQTM Sign [‘1’ for negative correlation between methylation and expression, ‘0’ otherwise], CpG DNAm [M-value], gene expression [log2(TPM+1)].

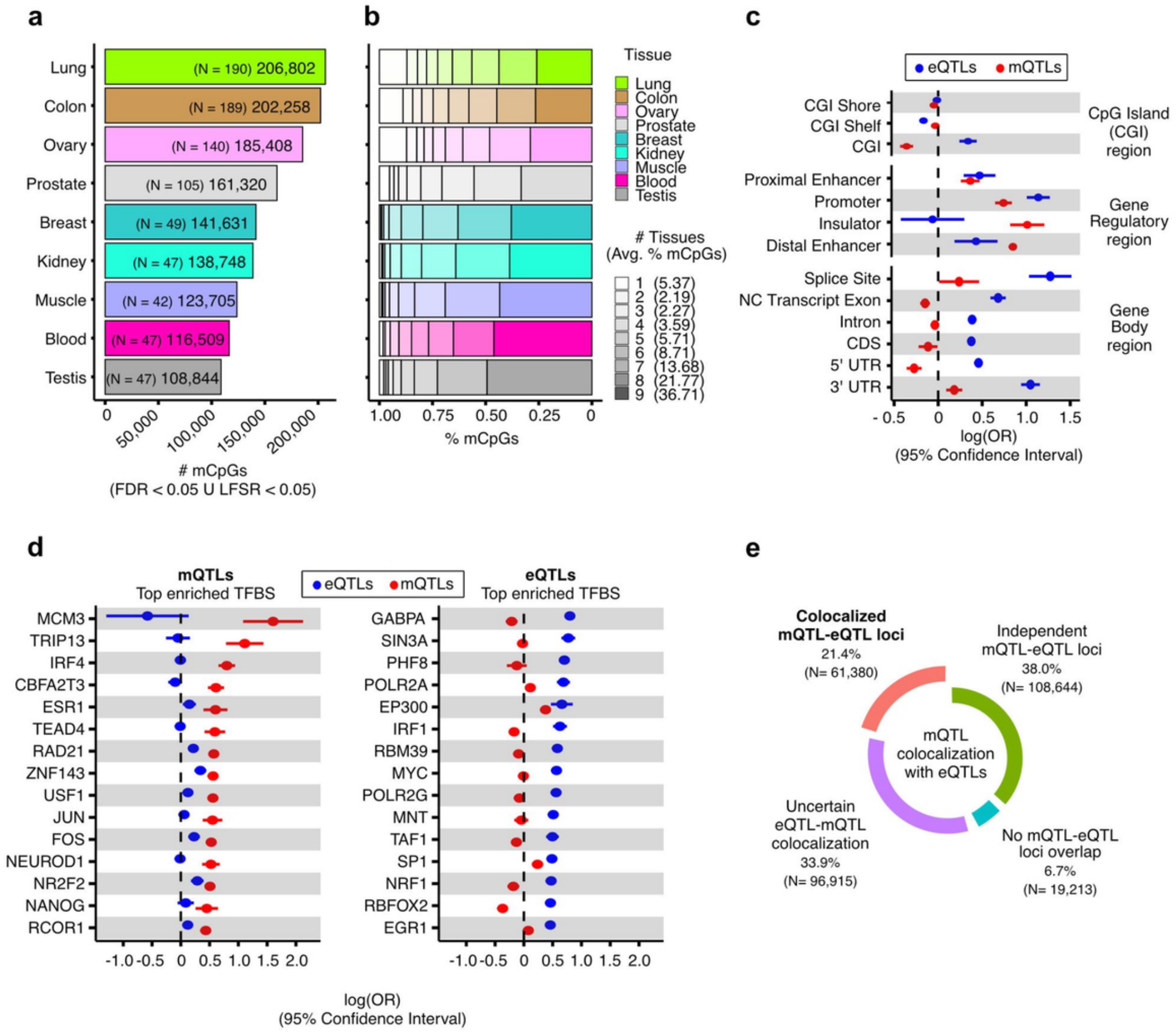

Figure 2

mQTL discovery and e/mQTL functional mechanism characterization. (a) Number of CpGs with a mQTL (mCpGs) per tissue, shown with per-tissue mQTL-mapping sample sizes in parentheses. (b) Cross-tissue sharing of mCpGs. Cross-tissue mean percent of mCpGs per tissue-sharing category is shown in parentheses. (c) QTL enrichment (x-axis) in CpG islands (CGI), gene body sites and candidate cis-regulatory elements. NC: Non-Coding. CDS: Coding Sequence. UTR: Untranslated Region. OR: Odds Ratio. (d) QTL enrichment (x axis) in transcription factor binding sites (TFBS). Top (largest OR value) 15 significant TFBS enrichments for mQTLs (left panel) and eQTLs (right panel) are shown. (e) Percent and number (in parentheses) of mQTL loci relative to eQTL-colocalization category.

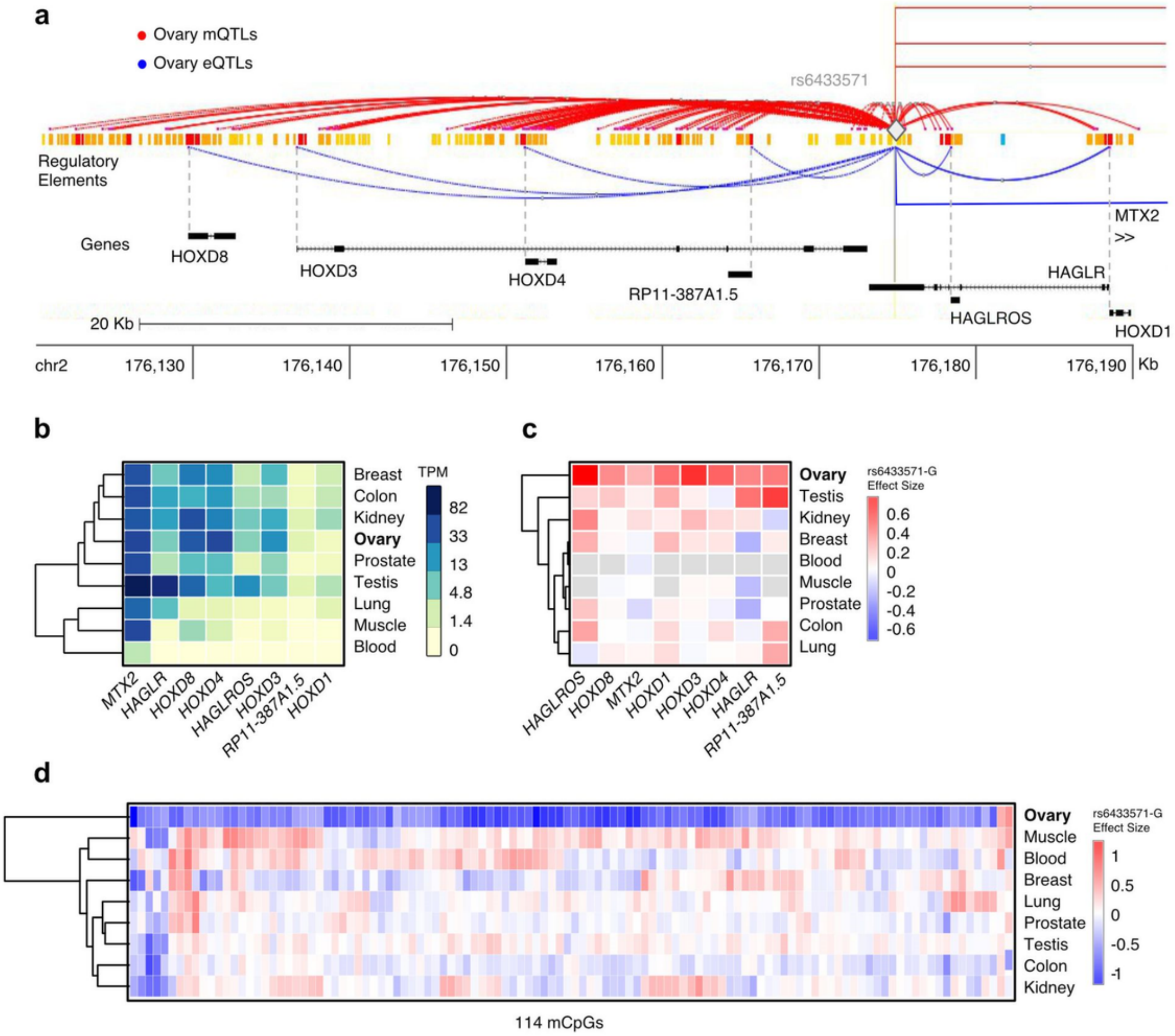

Figure 3

Characterization of HOXD-locus e/mQTL pleiotropy in the ovary. (a) Ovary-tissue mQTL and eQTL landscape for pleiotropic HOXD locus. Variant rs6433571 is depicted by a grey diamond and its QTL links with mCpGs and eGenes are depicted by red and blue arches, respectively. Rectangular arches indicate distal QTLs. Regulatory element annotations correspond to ENCODE candidate cis-Regulatory Elements (cCREs): promoters are shown in red, proximal and distal enhancers in orange and yellow, respectively, and CTCF-only regions (putative insulators) in blue. Dashed and solid grey lines correspond to gene transcription start sites and variant rs6433571 location, respectively. (b) Cross-tissue expression levels, in transcripts per million (TPM), of eGenes involved in the HOXD pleiotropic QTL locus identified in ovary tissue. (c-d) QTL effect size of variant rs6433571 minor allele G in identified HOXD-locus pleiotropic eGenes (c) and in mCpGs (d) per tissue. Grey-colored cells in (c) correspond to genes below expressed threshold in a particular tissue, hence not tested for eQTL signal. For panels (b-d), complete-linkage hierarchical clustering based on euclidean distance was performed for tissues and molecular phenotypes.

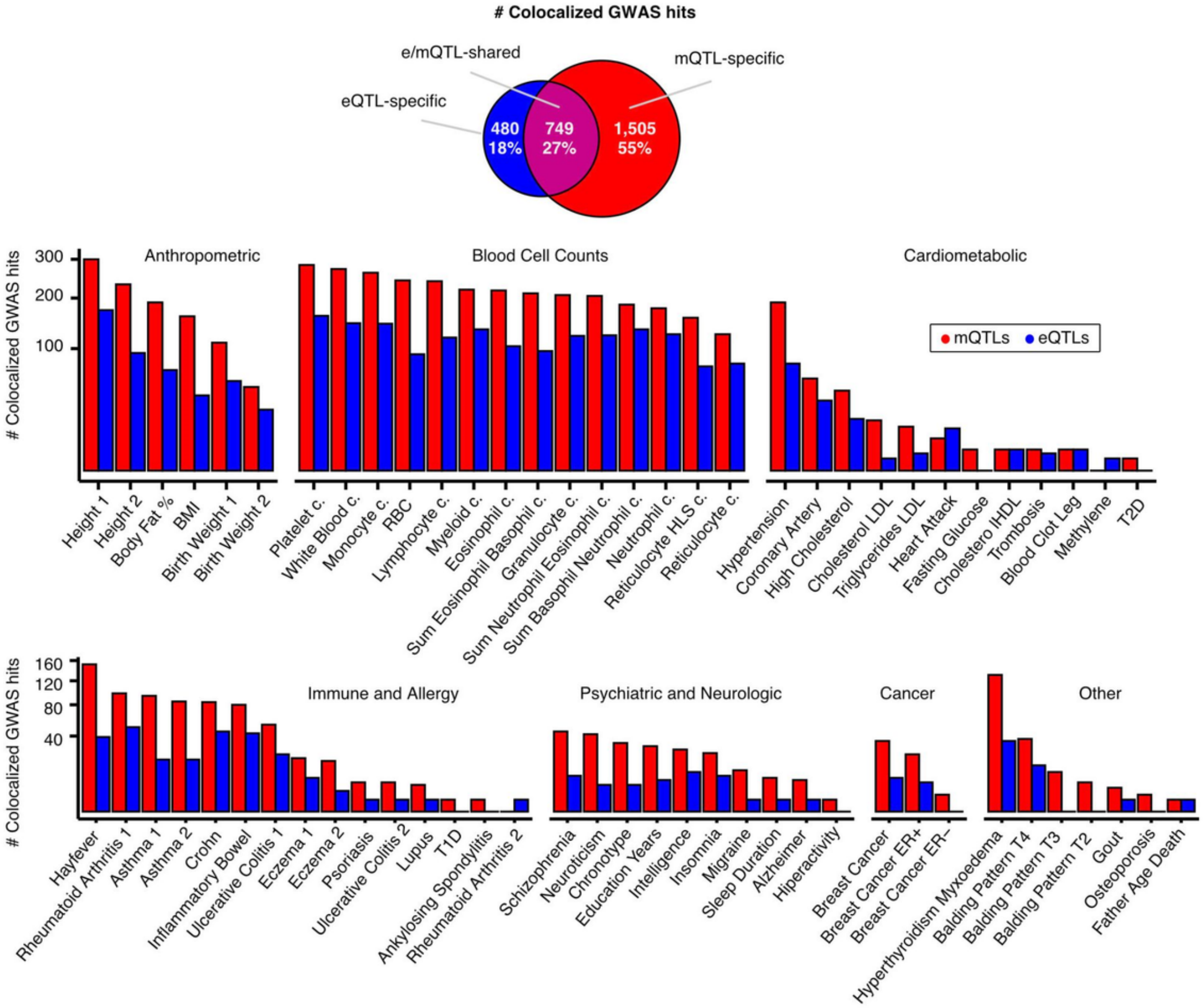

Figure 4

Colocalization of e/mQTLs with GWAS traits. Venn Diagram represents the overlap between colocalized GWAS hits per QTL type. Bar plot represents the number of colocalized GWAS hits (y axis, square-root transformed) per GWAS trait (x axis), trait class and QTL type. 'c.' stands for cell counts.

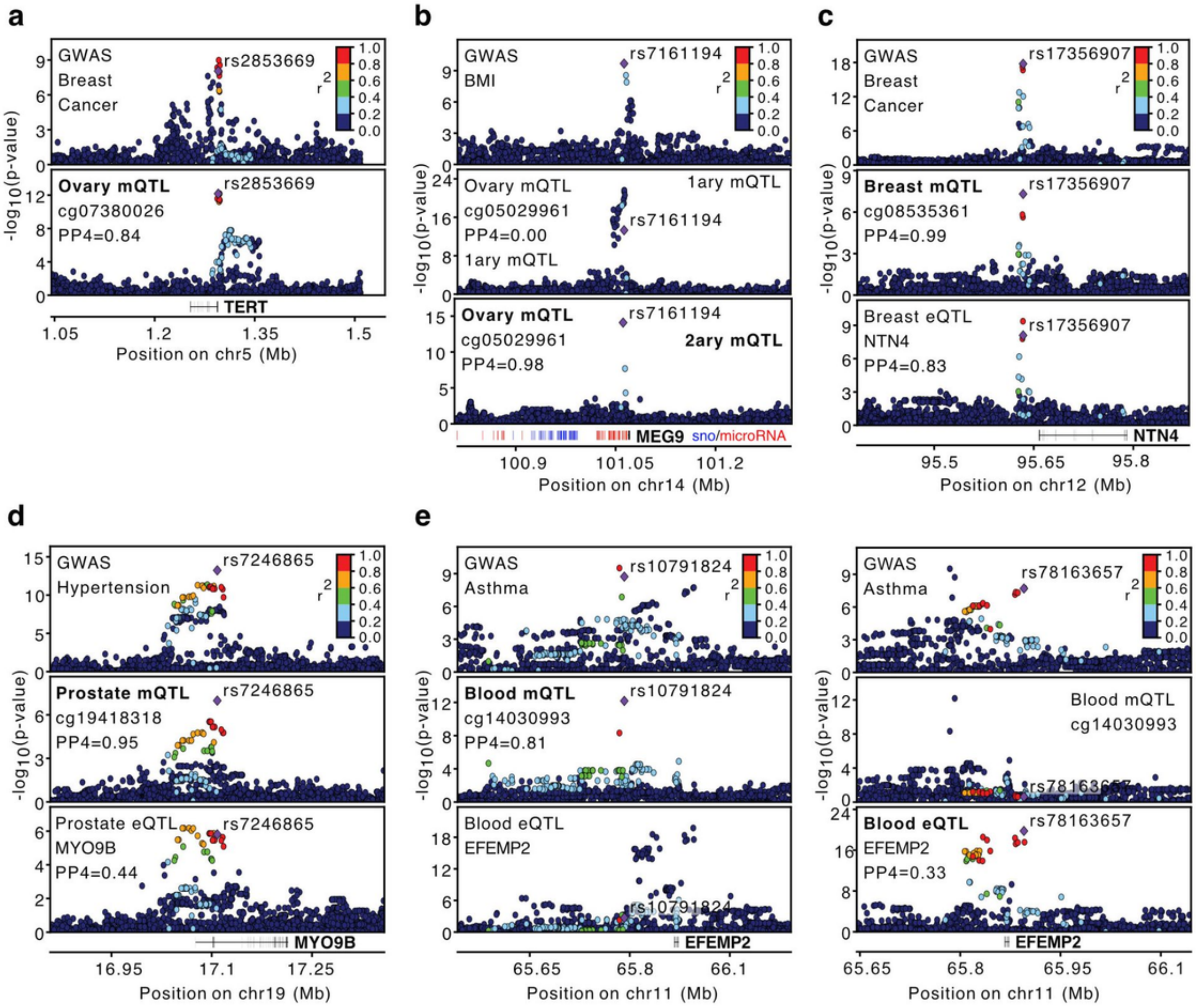

Figure 5

Trait-linked e/mQTLs. (a) Association p-values in the TERT locus for a breast cancer GWAS (top panel) and mQTL signal in ovary (bottom panel) (b) Genotype-phenotype association p-values of the MEG9 locus. Panels illustrate GWAS signal for body mass index (top), cg05029961 primary (middle) and secondary (bottom) mQTL signal in ovary tissue. Secondary mQTL association was obtained adjusting for the top significant variant of primary mQTL signal. Small nucleolar RNA loci (snoRNAs) and microRNA loci are depicted in blue and red, respectively. (c) Genotype-phenotype association p-values of the NTN4 locus. Panels illustrate GWAS signal for breast cancer (top), cg08535361 mQTL signal (middle) and NTN4 eQTL signal (bottom). (d) Genotype-phenotype association p-values of the MYO9B locus. Panels illustrate GWAS signal for hypertension (top), cg19418318 mQTL signal (middle) and MYO9B eQTL signal (bottom) in prostate. (e) Genotype-phenotype association p-values of the EFEMP2 locus. Top panels illustrate primary (left) and secondary (right) GWAS signals for asthma. Middle and bottom panels illustrate cg14030993 mQTL and EFEMP2 eQTL signal respectively. For all panels, top GWAS-colocalized e/mQTL is typed in bold, linkage disequilibrium between loci is quantified by squared Pearson coefficient of correlation (r2), and colocalization probability (PP4) of mQTL with GWAS signal is shown. In panels (a), (c) and (d), the diamond-shaped point represents the top significant mQTL variant, in panel (b) it represents the top significant secondary QTL variant, in panel (e) it represents either the top significant mQTL (left) or eQTL (right) variant.

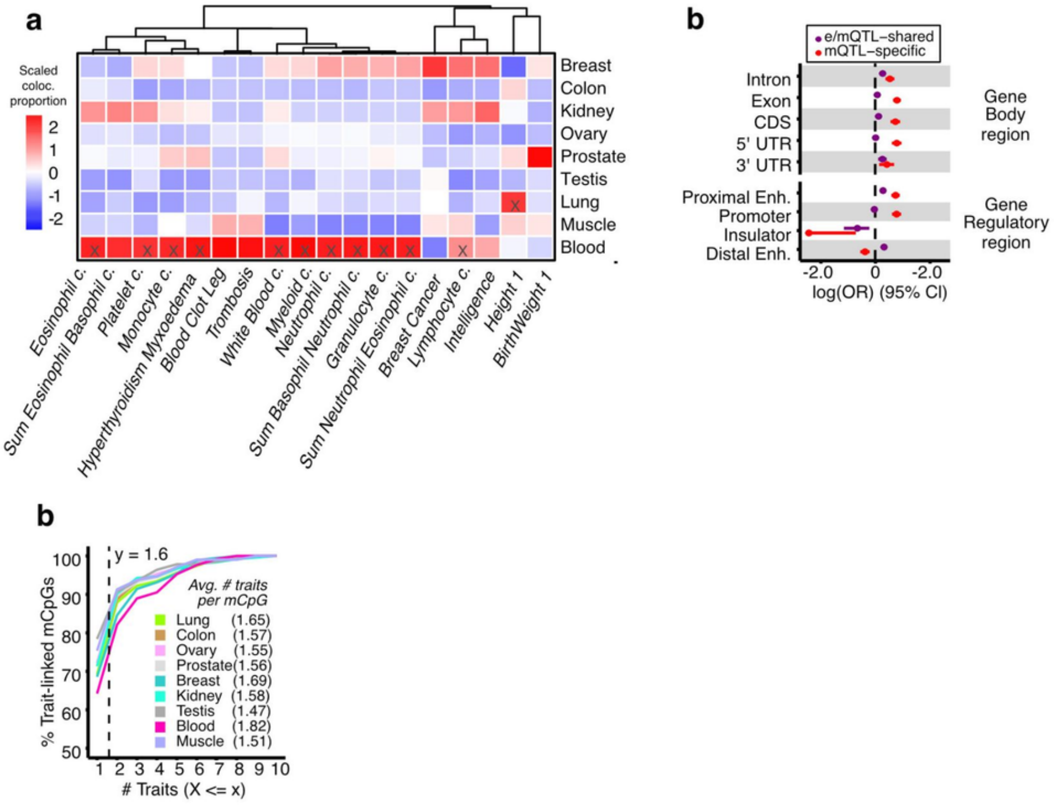

Figure 6

Characteristics of trait-linked mQTLs. (a) Proportion of colocalized mQTLs with GWAS hits per tissue (y axis) and GWAS trait (x axis) scaled by GWAS trait. GWAS traits are clustered by trait-wise scaled colocalization proportion values. Significant (Fisher's exact test FDR < 0.01) tissue-trait enrichments are labelled with a black cross. 'c.' stands for cell counts. 'coloc.' stands for colocalization. (b) Trait-linked mCpG enrichment (x-axis) in gene body regions and candidate cis-regulatory elements, stratified by QTL-GWAS colocalization group (see Fig. 4). CDS: Coding Sequence. UTR: Untranslated Region. OR: Odds Ratio.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SupplementaryTable1GTExv8cohortinformation.xlsx
- SupplementaryTable2eQTMfindings.xlsx
- SupplementaryTable3mQTLenrichmentinfunctionalannotations.xlsx
- SupplementaryTable4HOXDlocusQTLGWAScolocalization.xlsx
- SupplementaryTable5QTLGWAScolocalization.xlsx
- SupplementaryTable6TraitlinkedmCpGgenecandidates.xlsx
- SupplementaryMaterial.pdf